

Minimum-risk Sequence Alignment for the Alignment and Recognition of Action Videos

by
Zhen Wang

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

at the
School of Electrical and Data Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney

July 2018

Declaration of Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research was supported by an Australian Government Research Training Program (RTP).

Signed:	Production Note:
	Signature removed prior to publication.

Date:	18/07/2018
-------	------------

Abstract

Temporal alignment of videos is an important requirement of tasks such as video comparison, analysis and classification. In the context of action analysis and action recognition, the main guiding element for the temporal alignment are the human actions depicted in the videos. While well-established alignment algorithms such as dynamic time warping are available, they still heavily rely on basic linear cost models and heuristic parameter tuning. Inspired by the success of the hidden Markov support vector machine for pairwise alignment of protein sequences, in this thesis we present a novel framework which combines the flexibility of a pair hidden Markov model (PHMM) with the effective parameter training of the structural support vector machine (SSVM). The framework extends the scoring function of SSVM to capture the similarity of two input frame sequences and introduces suitable feature and loss functions. During learning, we leverage these loss functions for regularised empirical risk minimisation and effective parameter selection.

We have carried out extensive experiments with the proposed technique (nicknamed as EHMM-SSVM) against state-of-the-art algorithms such as dynamic time warping (DTW) and generalized canonical time warping (GCTW) on pairs of human actions from four well-known datasets. The results show that the proposed model has been able to outperform the compared algorithms by a large margin in terms of alignment accuracy.

In the second part of this thesis we employ our alignment approach to tackle the task of human action recognition in video. This task is highly challenging due to the substantial variations in motion performance, recording settings and inter-personal differences. Most current research focuses on the extraction of effective features and the design of suitable classifiers. Conversely, in this thesis we tackle this problem by a dissimilarity-based approach where classification is performed in terms of minimum distance from templates and where the distance is based on the score of our alignment model, the EHMM-SSVM. In turn, the templates are chosen by means of prototype selection techniques from the available samples of each class. Experimental results over two popular human action datasets

have showed that the proposed approach has been capable of achieving an accuracy higher than many existing methods and comparable to a state-of-the-art action classification algorithm.

Acknowledgements

I would like to take this opportunity to acknowledge the enormous support from my supervisor, Prof. Massimo Piccardi, the very useful suggestions from my group and the friendly environment of our lab especially during coffees, lunches and dinners.

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iv
List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 Background on temporal alignment	3
1.2 Minimum-risk temporal alignment	5
1.3 Classification based on alignment costs	5
1.4 Contributions	7
1.5 Publications	7
1.6 Thesis Outline	8
2 Review of Related Works	9
2.1 Dynamic Time Warping	10
2.2 Generalized Canonical Time Warping	13
2.2.1 Canonical Correlation Analysis (CCA)	14
2.2.2 Combination of CCA and DTW	16
2.3 Hidden Markov model	18
2.4 Pair hidden Markov model	20
2.5 Prototype Selection	23
2.6 Support vector machines	26
2.6.1 Binary SVM	27
2.6.2 Soft-margin SVM	30
2.7 Multi-Class SVM	31
2.7.1 One-vs-all classifier	32
2.7.2 One-vs-one classifier	33
2.7.3 Directed Acyclic Graph-support vector machines	34

2.8	The Structural Support Vector Machine	35
3	Pair Hidden Markov Support Vector Machine and its Application for the Alignment of Videos	40
3.1	Introduction	40
3.2	The Proposed Model: PHMM-SSVM	42
3.2.1	Pair Hidden Markov Model	42
3.2.2	Structural SVM	44
3.2.3	Integration	45
3.2.4	Parameter Vector and Feature Function	46
3.2.5	Loss Functions	47
3.2.6	Most-Violated Constraints	47
3.3	Experiments	50
3.3.1	Results on the Weizmann Dataset	51
3.3.2	Results on the Olympic Sports Dataset	52
3.4	Conclusion	53
4	The Extended Hidden Markov Support Vector Machine and its Application to the Alignment of Human Actions	54
4.1	Introduction	54
4.2	Extended Hidden Markov Model	55
4.3	The Proposed Model	57
4.3.1	Scoring Function	58
4.3.2	Dissimilarity Functions	59
4.3.3	Loss Function and Loss-Augmented Inference	59
4.3.4	Interpolation of the Ground-Truth Alignments Based on Key-Frames	61
4.4	Experimental Results	64
4.4.1	UCF101 Dataset	64
4.4.2	MSR Daily Activity 3D Dataset	65
4.5	Conclusion	67
5	Dissimilarity-Based Action Recognition with an Extended Hidden Markov Support Vector Machine	69
5.1	Introduction	69
5.2	EHMM-SSVM distance	71
5.3	Prototype selection with EHMM-SSVM	73
5.4	Classification	74
5.5	Experiments and Discussion	77
5.5.1	Results with the k -NN classifier on the KTH dataset	77
5.5.2	Results with the k -NN classifier on the Olympic Sports dataset . . .	79
5.5.3	Results with the DAGSVM classifier on both datasets	79
5.6	Conclusion	81
6	Conclusion	83

<i>Contents</i>	vii
-----------------	-----

Appendices	84
-------------------	-----------

Bibliography	85
---------------------	-----------

List of Figures

1.1	Example of the Euclidean distances between two sequences.	2
1.2	DTW distances between two sequences.	3
1.3	Example of optimal warping path produced by DTW.	4
1.4	Example of extended hidden Markov model. Latent variable y_i stores two paired indexes from sequences s and t	4
2.1	Illustration of paths of index pairs for some sequence s of length $N = 9$ and some sequence t of length $M = 7$. (A) Admissible warping path satisfying the conditions (i), (ii), and (iii) of Definition 2.1. (B) Boundary condition (i) is violated. (C) Monotonicity condition (ii) is violated. (D) Step size condition (iii) is violated.	11
2.2	Example of cost matrices in the DTW algorithm.	12
2.3	Hidden Markov model (HMM) graphical model.	18
2.4	PHMM state diagram.	22
2.5	A hyperplane separating two classes C1 and C2.	27
2.6	Geometric margins at each side of the hyperplane.	28
2.7	A single "outlier" can affect the separating hyperplane significantly.	30
2.8	Example of data from three hypothetical classes.	32
2.9	Examples of one-vs-all classifier.	32
2.10	Examples of one-vs-one classifier.	33
2.11	An example of DAGSVM	34
2.12	Illustration of a natural language parsing model. The graphics are reproduced from [37].	37
3.1	PHMM state diagram. Transitions "Insert a gap" between I_s and I_t are prohibited.	46
3.2	Sample actions	51
3.3	Example of ground-truth and predicted alignments from the Olympic Sports dataset: top two rows): six matched key-frames from two clean-and-jerk sequences. The first row is used as template and the frames of the second row show the ground-truth alignment; third row): alignment predicted by the proposed PHMM-SSVM; bottom two rows): alignments predicted by GCTW and DTW. The superimposed rectangles visually highlight the alignment errors.	52

4.1	The extended hidden Markov model for sequence alignment (represented as an undirected graphical model).	56
4.2	Main steps of the proposed approach: a) training (with regularized empirical risk minimization, i.e., structural SVM - Section 3.2.6); inference (Section 4.2).	56
4.3	Example of the interpolations on key-frames ground-truth alignments on the pairs of (1, 1), (5, 6), (9, 11), (12, 15) from UCF101 dataset.	62
4.4	Example of alignment obtained with the proposed approach for two videos of action "body weight squats" from UCF101.	66
5.1	Left: Example of frame-by-frame distances. Right: Example of dynamic time warping.	70
5.2	Examples of EHMM-SSVM alignment from the Olympic Sport dataset. Top two rows: 14 matches and inserted gaps for two paired "clean-and-jerk" sequences. Bottom two rows: 16 matches and inserted gaps for two paired "tennis-serve" sequences (the gaps are inserted only on the second sequence because of its slower execution).	71
5.3	The EHMM-SSVM as a graphical model.	72
5.4	Sample actions in the Olympic dataset (top four rows) and KTH dataset (bottom row).	76
5.5	The EHMM-SSVM distance matrix between action instances in KTH dataset. Darker colours denote higher similarity.	78

List of Tables

2.1	Numerical example of the cost matrix and the accumulated cost matrix determined during the formation of the optimal warping path. (A) Each cell represents the distance between elements s_i and t_j and is computed as cost function $C(s_i, t_j) = s_i - t_j $. (B) The accumulated cost matrix is obtained using Definition 2.2.	13
2.2	Transition probabilities table.	22
3.1	Transition probabilities table.	46
3.2	Alignment accuracy for action “jump” in the Weizmann dataset.	51
3.3	Alignment accuracy for action “clean-and-jerk” in the Olympic Sports dataset.	53
4.1	Transition probabilities table.	56
4.2	Experimental results for action “body weight squats” from the UCF101 dataset.	65
4.3	Experimental results for action “stand up” from the MSR Daily Activity 3D dataset.	65
4.4	Example data from the MSR Daily Activity 3D dataset.	67
4.6	Example of skeleton alignment from action “stand up” in the MSR Daily Activity 3D dataset. The arrows highlight noticeable inaccuracies.	68
5.1	Example of EHMM-SSVM distances between a <i>walking</i> sample in the KTH dataset and prototypes from the various classes. Darker colours denote higher similarity.	76
5.2	Accuracy with the k -NN classifier on the KTH dataset.	79
5.3	Accuracy in the second experiment on the Olympic Sports dataset.	80
5.4	Accuracy with the DAGSVM classifier on the KTH dataset.	80
5.5	Accuracy with the DAGSVM classifier on the Olympic Sports dataset.	81